Automatic Query Type Identification Based on Click Through Information

Yiqun Liu[1], Min Zhang[1], Liyun Ru[2], and Shaoping Ma[1]

[1] State Key Lab of Intelligent Tech. & Sys., Tsinghua University, Beijing, China
liuyiqun03@mails.tsinghua.edu.cn
[2] Sogou Incorporation, Beijing, China
ruliyun@sohu-rd.com

Abstract. We report on a study that was undertaken to better identify users' goals behind web search queries by using click through data. Based on user logs which contain over 80 million queries and corresponding click through data, we found that query type identification benefits from click through data analysis; while anchor text information may not be so useful because it is only accessible for a small part (about 16%) of practical user queries. We also proposed two novel features extracted from click through data and a decision tree based classification algorithm for identifying user queries. Our experimental evaluation shows that this algorithm can correctly identify the goals for about 80% web search queries.[1]

1 Introduction

Web Search engine is currently one of the most important information access and management tools for WWW users. Most users interact with search engine using short queries which are composed of 4 words or even fewer. This phenomena of "short queries" has prevented search engines from finding users' information needs behind their queries.

With analysis into search engine user behavior, Broder [1] and Rose [2] independently found that search goals behind user queries can be informational, navigational or transactional (refered to as resource type by Rose). Further experiment results in TREC [3][4] showed that informational and navigational search results benefit from different kinds of evidences. Craswell [5] and Kraaij [6] found that anchor text and URL format offer improvement to content-only method for home page finding task, which covers a major part of navigational type queries. Bharat [7] proved that informational type searches may be improved using hyper link structure analysis. According to these researches, if query type can be identified for a given user query, retrieval algorithm can be adapted to this query type and search performance can be improved compared with a general purpose algorithm. That is why we should identify users' search goals behind their submitted queries.

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60223004, 60321002, 60303005, 60503064) and the Key Project of Chinese Ministry of Education (No. 104236).

H.T. Ng et al. (Eds.): AIRS 2006, LNCS 4182, pp. 593–600, 2006.

Query type identification can be performed by two means: Sometimes queries can be classified by simply looking at its content. "AIRS2006" is a navigational type query because the user probably wants to find the homepage of this conference; while "car accident" may be informational because the user seems to be looking for detailed information on "car accident". However, several queries can only be classified with the help of search context. For the query "information retrieval", it is impossible to guess without further information whether the user wants to locate the book written by CJ van Rijsbergen (navigational) or he wants to know something about IR (informational).

The remaining part of the paper is constructed as follows: Section 2 analyzes into search engine logs and studies the possibilities of using click through data and anchor text in query type prediction. Section 3 proposed two novel features extracted from click through data and developed a decision tree based classification algorithm. Experimental results of query classification are shown in Section 4. Finally come discussion and conclusion.

2 Analysis into Search Engine Logs

In order to verify reliability and scalability of our classification method, we obtained part of query logs from a widely-used Chinese search engine Sogou (www.sogou.com). The logs are collected from February 1st to February 28th in the year 2006. They contain 86538613 non-empty queries and 4345557 of them are unique. Query sessions are provided according to cookie information and there are totaly 26255952 sessions in these logs.

When we try to predict one user query's type, we can make use of click through data if and only if this query appears in past click through logs. In Figure 1, the category axis shows date when the logs are collected (all logs are collected in February 2006, so year/month information is omitted) and the value axis show the percentage of queries which have click through information.

We can see that new queries made up of 100% queries on the first day because no history information is available before that day. However, as time goes by the percentage of newly-appeared queries drops to about 10% (average data is 11.15% for the last 10 days). It means that click through data can be applied to classify about 90% queries for search engines.

According to previous works [8] and [9], if one web page shares the same anchor text as a query, the query is probably navigational type. In those works which use anchor text for query type identification, only those queries which match a certain number of anchor texts can be predicted using this evidence. So it is important to find how many percentage of web search queries have such matches with anchor texts.

We crawled over 202M Chinese web pages from the Web and extracted anchor text information from these pages. After reducing possible spams, noises and redundancies, the percentage of queries matching a certain number of anchor text is calculated. We found that the percentage of matching queries doesn't vary with time and there are less than 20% (16.24% on average) matching queries

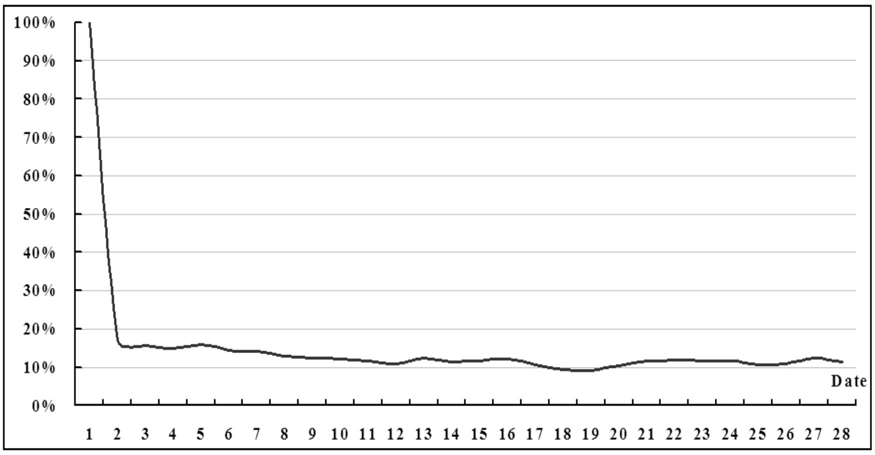

Fig. 1. Percentage of newly-appeared queries in query logs

each day. It means that anchor text evidence is only applicable for less than 20% queries in practical Web search environment.

We can conclude that click through data is suitable for query type identification for general purpose Web search engine. Anchor text evidence may be effective for a part of queries but it is not applicable for the major part. So our work is mainly focused in classification using click through evidence.

3 Query Type Identification Using Click Through Data

In this section, we propose two novel evidences extracted from click through data: *n Clicks Satisfied (nCS)* and *top n Results Satisfied (nRS)*. They can be used as features in our query type identification algorithm.

In order to find the differences between navigational and informational / transactional type queries, we developed a training set of queries which contains 153 navigational queries and 45 informational queries. These queries are randomly selected from query logs and manually classified by 3 assessors using voting to decide queries' categories.

3.1 N Clicks Satisfied (nCS) Evidence

N Clicks Satisfied (nCS) evidence is extracted from the number of user clicks for a particular query. It is based on the following assumption:

Assumption 1 (Less Effort Assumption): While performing a navigational type search request, user tend to click a small number of URLs in the result list.

Supposing one web search user has a navigational goal (CNN homepage, for example), he has a fixed search target in mind and would like to find just that target URL (www.cnn.com) and corresponding snippet in the result list. So it

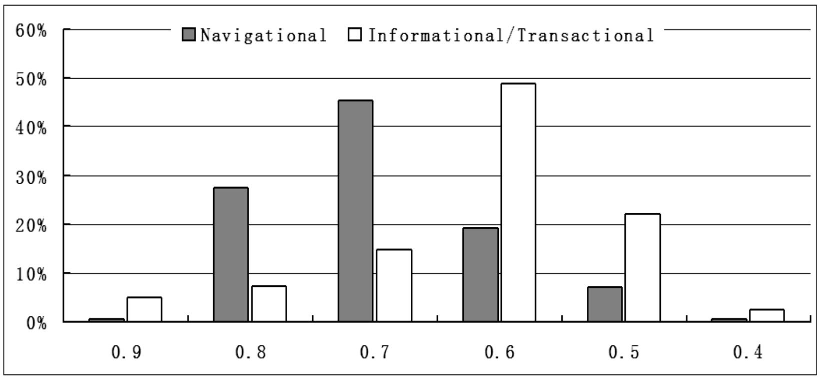

Fig. 2. nCS feature distribution in the training set when $\boldsymbol{n}$ is set to 2. Category axis shows nCS value and value axis shows the percentage of navigational/informational/transactional queries with a certain nCS value.

is impossible for him to click a number of URLs which are not the target page unless there exists cheating pages.

According to the *Less Effort Assumption*, we can judge a query type by the number of URLs which the user clicks. nCS feature is defined as:

$$nCS(Query\ \boldsymbol{q}) = \frac{\#(Session\ of\ \boldsymbol{q}\ that\ involves\ less\ than\ \boldsymbol{n}\ clicks)}{\#(Session\ of\ \boldsymbol{q})}. \tag{1}$$

According Figure 2, navigational type queries have larger *nCS* than informational/transactional ones. Most navigational queries have a nCS larger than 0.7 while 70% informational/transactional queries' nCS is less than 0.7. It means this feature can separate a large part of navigational queries.

3.2 Top n Results Satisfied (nRS) Evidence

Top n Results Satisfied (nRS) evidence is extracted from the clicked URL's rank information. It is based on the following assumption:

Assumption 2 (Cover Page Assumption): While performing a navigational type search request, user tend to click only the first few URLs in the result list.

This assumption is related to the fact that navigational type queries have a much higher retrieval performance than informational/transactional ones. According to TREC web track and terabyte track experiments [3],[4] and [10] in the last few years, an ordinary IR system can return correct answers at the 1st ranking for 80% user queries.

According to the *Less Effort Assumption*, we can judge a query type by whether the user clicks other URLs besides the first $\boldsymbol{n}$ ones. Top n Results Satisfied (nRS) feature developed from this idea is defined as:

$$nRS(Query\ \boldsymbol{q}) = \frac{\#(Session\ of\ \boldsymbol{q}\ that\ involves\ clicks\ only\ on\ top\ \boldsymbol{n}\ results)}{\#(Session\ of\ \boldsymbol{q})}. \tag{2}$$

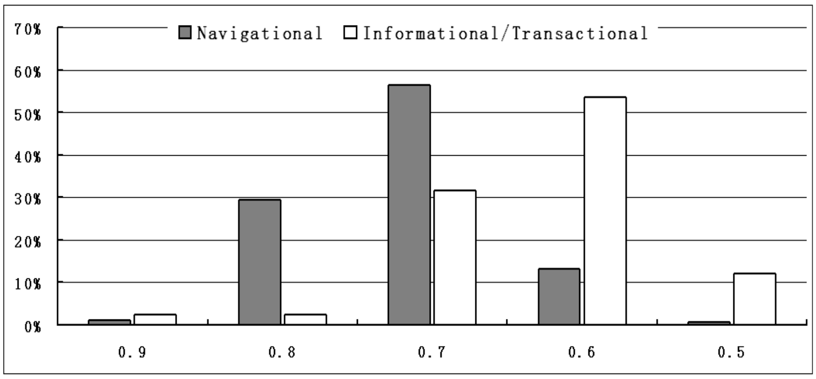

Fig. 3. nRS feature distribution in the training set when n is set to 5. Category axis shows nRS value and value axis shows the percentage of navigational/informational/transactional queries with a certain nRS value.

According to the nRS distribution shown in Figure 3, navigational type queries have larger nRS than informational/transactional ones. 80% navigational queries have a nRS value larger than 0.7 while about 70% informational/transactional queries' nRS is less than 0.7. It shows this feature can also be used to classify web search queries as well as nCS.

3.3 A Learning Based Identification Algorithm

Based on the two new features proposed in Section 4.1 and 4.2, we can separate informational/transactional queries from navigational ones. Besides these features, click distribution proposed by Lee [8] is also believed to be able to identify web search queries.

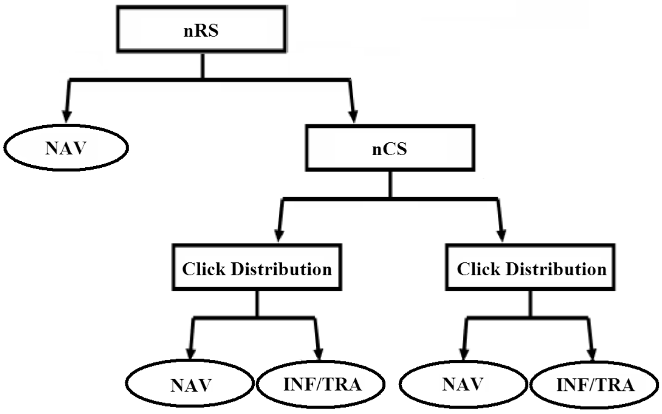

Fig. 4. A Query identification decision tree composed of click through features(INF:informational, TRA:transactional, NAV:navigational)

In order to combine these 3 features: nCS, nRS and click distribution to finish the query type identification task, we adopted a typical decision tree algorithm. It is a method for approximating discrete-valued functions that is robust of noisy data and capable of learning disjunctive expressions. We choose decision tree because it is usually the most effective and efficient classifier when we have a small number (3 features here) of features.

We used standard C4.5 algorithm to combine these 3 features and get the following decision tree shown in Figure 4. According to C4.5 algorithm, The effectiveness of features can be estimated by the distance away from the root. We can see that nRS is more effective in identification than nCS and click distribution. The two new features proposed are more reliable here according to the metric of information ratio in C4.5 algorithm.

4 Experiments and Discussions

4.1 Query Type Identification Test Set and Evaluation Measures

We developed a test set to verify the effectiveness of our identification algorithm. This test set is composed of 81 informational/transactional type queries and 152 navigational queries. The informational/transactional ones were obtained from a Chinese search engine contest organized by tianwang.com (part of the contest is specially designed to test the search engines' performance for informational/transactional queries) and the navigational ones are obtained from a widely-used Chinese web directory hao123.com (websites and their description are used as navigational type results and queries correspondingly). This test set is not assessed by the people who developed the training set in order to get rid of possible subjective noises.

We use traditional precision/recall framework to judge the effectiveness of the query type identification task. Precision and Recall values are calculated separately for two kinds of queries. They are also combined to F-measure value to judge the overall performance.

4.2 Query Type Identification Results

The test query set described in the previous section are identified by our decision tree algorithm. Experiment results are shown in Table 1. We also compared our method with Lee's Click-Distribution method [8] in Figure 5 because it is to our knowledge the most effective click-through information based method.

According to the experimental results in Table 1, precision and recall values over 80% are achieved to identify web search queries. It shows that most web search queries are successfully identified with the help of click through information. We further found that our decision tree based method outperforms Click-Distribution method which is proposed by Lee[8] both on the training set (improved by 21%) and on the test set(improved by 18%). Although Click-Distribution is quite effective for query type identification, by adding new features of nCS and nRS, better performance is achieved.

Table 1. Query type identification experimental results.(INF:informational, TRA:transactional, NAV:navigational)

	Training set			Test set		
	INF/TRA	NAV	Mixed	INF/TRA	NAV	Mixed
Precision	76.00%	91.07%	87.65%	73.74%	85.62%	81.49%
Recall	66.67%	90.71%	85.25%	72.84%	86.18%	81.54%
F-measure	0.71	0.91	0.86	0.73	0.85	0.81

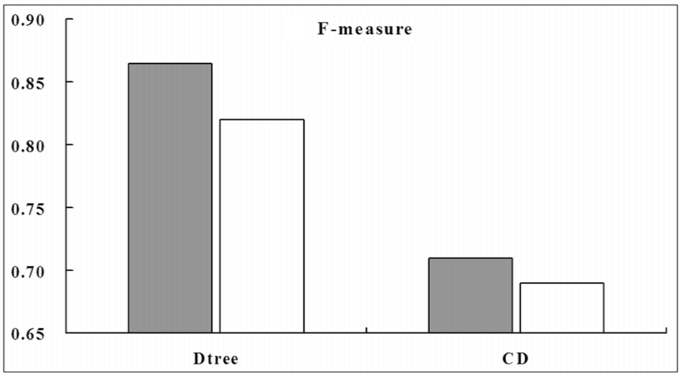

Fig. 5. F-measure values of our decision tree based method and the Click-Distribution based method, Dtree: decision tree based method, CD: Click-Distribution based method, Train/Test: experimental results based on the training/test set

5 Conclusions and Future Work

Given the vast amount of information on the World Wide Web, a typical short query of 1-3 words submitted to a search engine usually cannot offer enough information for the ranking algorithm to give a high quality result list. Using click through data to identify the user goals behind their queries may help search engine to better understand what users want so that more effective result ranking can be expected.

Future study will focus on following aspects: How well does these new features work together with evidences from the queries themselves? How should the traditional ranking models be adjusted for automatically identified queries? Is there any other proper query classification criterion in users' notion?

References

1. Andrei Broder: A taxonomy of web search. SIGIR Forum(2002), Volume 36(2):3-10.
2. Daniel E. Rose and Danny Levinson, Understanding User Goals in Web Search. In proceedings of the 13th World-Wide Web Conference, 2004.

3. N. Craswell, D. Hawking: Overview of the TREC-2002 web track. In The eleventh Text Retrieval Conference (TREC-2002), NIST, 2003.
4. N. Craswell, D. Hawking: Overview of the TREC-2003 web track. In the twelfth Text REtrieval Conference (TREC 2003), NIST, 2004.
5. N Craswell, D Hawking and S Robertson. Effective Site Finding using Link Anchor Information. In Proceedings of ACM SIGIR '01, 2001.
6. Kraaij W, Westerveld T, Hiemstra D. The importance of prior probabilities for entry page search. In Proceedings of ACM SIGIR '02, 2002.
7. Bharat K, Henzinger M. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of ACM SIGIR '98, 1998.
8. Uichin Lee, Zhenyu Liu and Junghoo Cho, Automatic Identification of User Goals in Web Search, in proceedings of the 14th World-Wide Web Conference, 2005.
9. I. Kang and G. Kim. Query type classication for web document retrieval. In Proceedings of ACM SIGIR '03, 2003.
10. N. Craswell and D. Hawking. Overview of the TREC-2004 Web track. In the Thirteenth Text REtrieval Conference Proceedings (TREC 2004), NIST, 2005.